Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

August 19, 2019

Topics

Objectives

Objectives:

- Differentiate between a population parameter and a sample statistic.
- Compute and interpret the sample mean of a data set.
- Compute and interpret the sample median of a data set.
- Know which statistic (mean or median) is resistant to outliers.

Summarizing Quantitative Data: Measures of Center

(3.1)

Statistics and Population Parameters

- A ***statistic*** is a numerical value computed from data in a *random sample*.

Summarizing Quantitative Data: Measures of Center

(3.1)

Statistics and Population Parameters

- A ***statistic*** is a numerical value computed from data in a *random sample*.
- A ***parameter*** is a numerical value computed from data over the *entire population*.

- We'll look at two statistics that measure the center (or "typical value") of a data set:
  1. The sample mean
  2. The sample median

- We'll look at two statistics that measure the center (or "typical value") of a data set:
  1. The sample mean
  2. The sample median
- **We'll denote the sample observations (data) as $x_1, x_2, x_3, \ldots, x_n$, where n is the sample size.**

The Sample Mean

- To compute the ***sample mean***, denoted $\bar{x}$:

Sample Mean:

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum x_i \end{aligned}$$

where $\sum$ is the so-called ***summation symbol*** , with $\sum x_i$ telling us to add the x_i's together.

The Sample Mean

- To compute the ***sample mean***, denoted $\bar{x}$:

Sample Mean:

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum x_i \end{aligned}$$

where $\sum$ is the so-called ***summation symbol*** , with $\sum x_i$ telling us to add the x_i's together.

- **The sample mean is just the *average*** of the observations in the sample.

Example

A person's metabolic rate is the rate at which the body consumes energy while at rest. The data below are the metabolic rates (in calories per day) for 7 men who took part in a study of dieting and weight loss.

1792 1666 1362 1614 1460 1867 1439

Example

A person's metabolic rate is the rate at which the body consumes energy while at rest. The data below are the metabolic rates (in calories per day) for 7 men who took part in a study of dieting and weight loss.

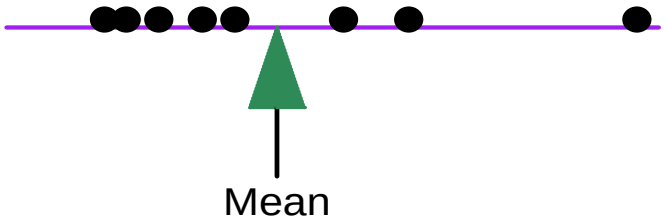
1792 1666 1362 1614 1460 1867 1439

The sample **mean** is

$$\begin{aligned} \bar{x} &= \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} \\ &= 1600. \end{aligned}$$

- If the data were weights along a (weightless) horizontal axis, the weights would **balance** at $\bar{x}$.

The Mean as the Balancing Point of a Data Set

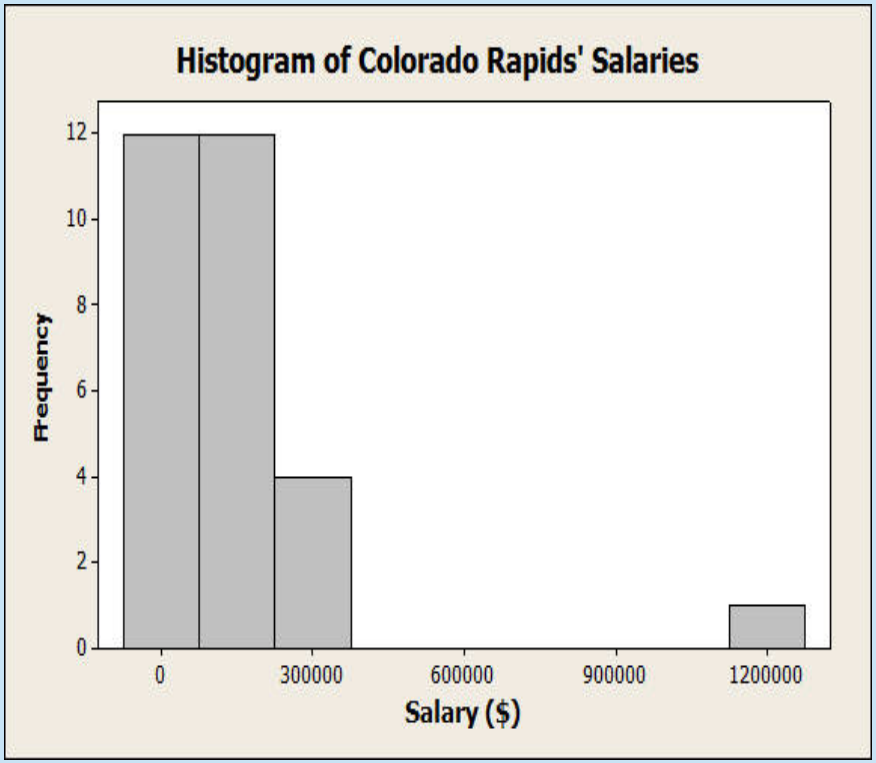


- It may be the case that **more than half** of the data lie **below** (or **above**) the **mean**, as in the next example.

Example

The table on the next slide shows the salaries of players on the 2015 Colorado Rapids professional soccer team.

Player	Salary
Carlos Alvarez (M)	$70,000
Dominique Badji (F)	$50,000
John Berner (GK)	$60,000
Marc Burch (D)	$110,000
Bobby Burling (D)	$132,500
Caleb Calvert (F)	$65,000
Sam Cronin (M)	$200,000
Kevin Doyle (F)	$1,125,000
Charles Eloundou (F)	$60,000
Joseph Greenspan (D)	$50,000
Marlon Hairston (M)	$80,000
Michael Harrington (D)	$130,000
Clint Irwin (GK)	$85,000
Nick LaBrocca (M)	$160,000
Zac MacMath (GK)	$130,000
Drew Moor (D)	$258,500
Ben Newnam (D)	$60,000
Shane O'Neill (M)	$64,998
Lucas Pittinari (M)	$190,000
Dillon Powers (M)	$245,000
Juan Edgardo Ramirez (M)	$75,000
James Riley (D)	$77,500
Vicente Sanchez (F)	$210,000
Marcelo Sarvas (M)	$360,000
Dillon Serna (M)	$60,000
Axel Sjoberg (D)	$60,000
Luis Solignac (F)	$65,004
Gabriel Torres (F)	$262,000
Jared Watts (M)	$60,000



The **mean** salary is **$157,086**. But a quick glance through the data reveals that **more than half** (20 of the 29 players) earn **less than the mean**.

The Population Mean

- The ***population mean***, denoted μ, is the average of the x's in the entire population To compute the population mean:

Population Mean:

$$\mu = \frac{1}{N}\sum x_i$$

where N is the population size.

The Sample Median

- To compute the ***sample median***: First **sort** the data from smallest to largest. Then

Sample Median:

$$\text{Sample Median} = \begin{cases} \text{The middle value if } n \text{ is odd} \\ \text{The average of the two middle values if } n \text{ is even} \end{cases}$$

The Sample Median

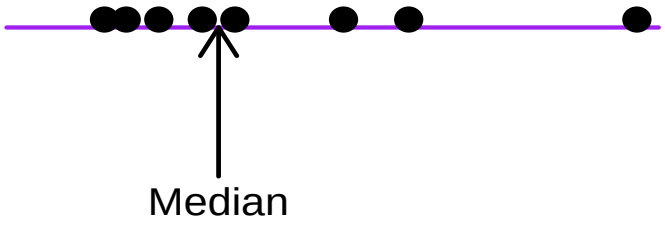
- To compute the ***sample median***: First **sort** the data from smallest to largest. Then

Sample Median:

$$\text{Sample Median} = \begin{cases} \text{The middle value if } n \text{ is odd} \\ \text{The average of the two middle values if } n \text{ is even} \end{cases}$$

- When n is even, **exactly half** of the data will be **less** than the **median** and the other half **greater**. When n is odd, it's approximately half of the data.

The Median as the Midpoint of a Data Set



- The next example shows how to find the **median** when the sample size is *odd*.

Example

Here are the metabolic rates **sorted** from smallest to largest.

1362 1439 1460 1614 1666 1792 1867

- The next example shows how to find the **median** when the sample size is *odd*.

Example

Here are the metabolic rates **sorted** from smallest to largest.

$$1362 \quad 1439 \quad 1460 \quad 1614 \quad 1666 \quad 1792 \quad 1867$$

Because $n = 7$ is **odd**, the sample **median** is the middle value,

$$\text{Median} \ = \ 1614.$$

- The next example shows how to find the **median** when the sample size is *even*.

Example

The following data represent lifetimes (in hours) of a sample of $n = 10$ incandescent light bulbs:

898 964 970 983 1003 1016 1022 1029 1058 1085

(Note that the data are already **sorted**).

- The next example shows how to find the **median** when the sample size is *even*.

Example

The following data represent lifetimes (in hours) of a sample of $n = 10$ incandescent light bulbs:

898 964 970 983 1003 1016 1022 1029 1058 1085

(Note that the data are already **sorted**).

Because $n = 10$ is **even**, the sample **median** is the average of the two middle values,

$$\text{Median} = \frac{1003 + 1016}{2} = 1009.5.$$

Computing the Mean from Grouped Data

- The following example illustrates that the mean of a set of data can be computed even when the data are **grouped** (as in a frequency distribution table).

Example

A certain international non-governmental organization has three employee levels (entry level, mid career, and senior level) with monthly salaries \$3,500, \$4,500, and \$6,000, respectively.

Here are the monthly salaries of $n = 10$ employees at the organization.

3,500	3,500	4,500	3,500	3,500	3,500	6,000	4,500
4,500	3,500						

The **mean** of the data is:

$$\begin{aligned} \bar{x} &= \frac{1}{10}(3,500 + 3,500 + 4,500 + 3,500 + 3,500 + 3,500 \\ &\quad +6,000 + 4,500 + 4,500 + 3,500) \end{aligned}$$

We can **group** the six 3,500's together, the three 4,500's together, and the one 6,000 with itself, to get:

$$\begin{aligned} \bar{x} &= \frac{1}{10}\left(3,500 \times 6 + 4,500 \times 3 + 6,000 \times 1\right) \\ &= \frac{1}{10}\left(40,500\right) \\ &= 4,050. \end{aligned}$$

This suggests a way to compute a **mean** directly from a frequency distribution table:

Salary	Frequency
3,500	6
4,500	3
6,000	1
	$n = 10$

This suggests a way to compute a **mean** directly from a frequency distribution table:

Salary	Frequency
3,500	6
4,500	3
6,000	1
	$n = 10$

In particular, we can compute the **mean** by multiplying each salary by its frequency, summing the results, and dividing that sum by the total number of employees.

- The previous example illustrates the idea behind the following formula, which indicates how to compute a mean directly from a frequency distribution table:

Sample Mean (Grouped Data Formula):

$$\bar{x} = \frac{1}{n}\sum x_i f_i$$

where here each x_i is one of the **unique** values in the data set and f_i is the **frequency** of that value.

Resistance to Outliers

- We say that a statistic is ***resistant*** to outliers if its value is not strongly influenced by their presence in the data.

Resistance to Outliers

- We say that a statistic is ***resistant*** to outliers if its value is not strongly influenced by their presence in the data.
- **The sample median is resistant, but the sample mean is *not* resistant.**

Exercise

Suppose the largest metabolic rate in the earlier example was 3000 calories per day (an **outlier**) instead of 1867:

1362 1439 1460 1614 1666 1792 ~~1867~~ 3000

Exercise

Suppose the largest metabolic rate in the earlier example was 3000 calories per day (an **outlier**) instead of 1867:

1362 1439 1460 1614 1666 1792 ~~1867~~ 3000

Recompute the **mean** and **median** for this new set of data, and note how each statistic is (or isn't) influenced by the presence of the **outlier**.

Approximating the Mean and Median from a Histogram

- We can approximate the values of the sample mean and sample median by looking at a histogram of the data:

Approximating the Mean and Median from a Histogram

- We can approximate the values of the sample mean and sample median by looking at a histogram of the data:
  - The **mean** is the point on the horizontal axis at which the bars would balance if they were weights (i.e. the **"balancing point"** of the histogram).

Approximating the Mean and Median from a Histogram

- We can approximate the values of the sample mean and sample median by looking at a histogram of the data:
  - The **mean** is the point on the horizontal axis at which the bars would balance if they were weights (i.e. the **"balancing point"** of the histogram).
  - The **median** is point on the horizontal axis for which 50% of the *area* of the bars lies to either side (i.e. the **"equal areas point"** of the histogram).

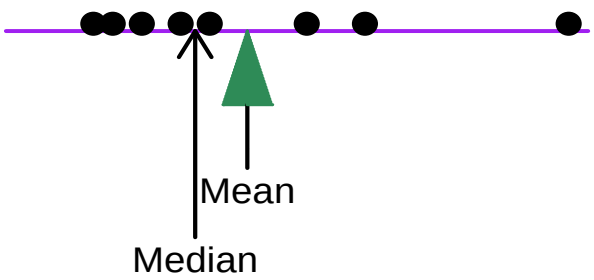
- Note that:
  - **For a *right skewed* histogram, the mean will be *greater* than the median.**

- Note that:
  - **For a *right skewed* histogram, the mean will be *greater* than the median.**
  - **For a *left skewed* histogram, the mean will be *smaller* than the median.**

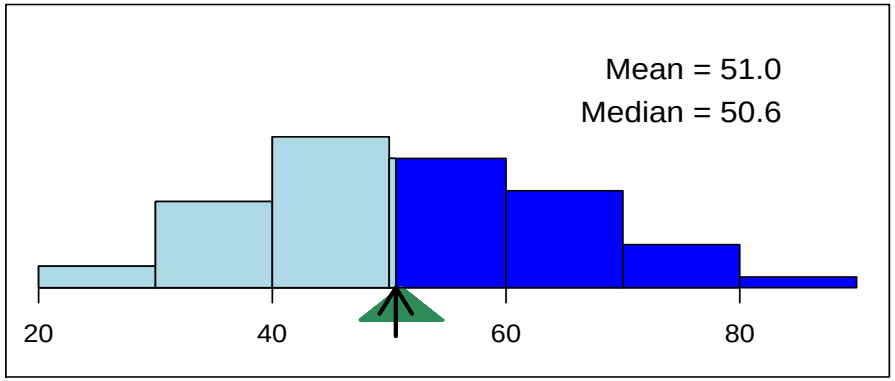
- Note that:
  - **For a *right skewed* histogram, the mean will be *greater* than the median.**
  - **For a *left skewed* histogram, the mean will be *smaller* than the median.**
  - **For a histogram that's roughly *symmetric*, the mean and median will be approximately *equal*.**

The Mean as the Balancing Point of a Data Set

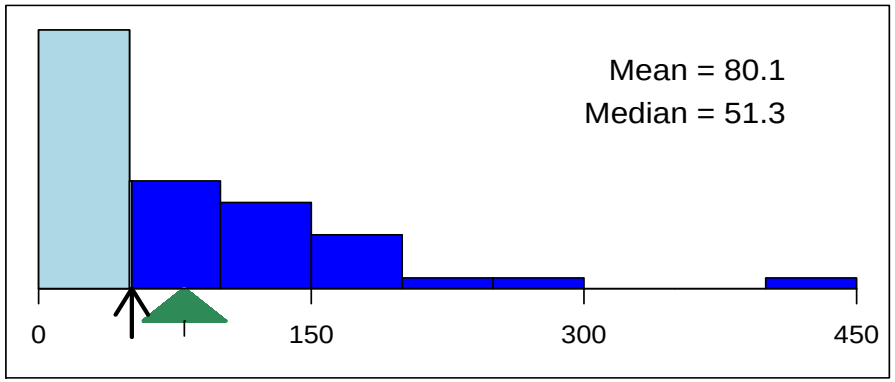
and the Median as the Midpoint



Symmetric Distribution, Mean ≈ Median



Right Skewed Distribution, Mean > Median



Exercise

We can think of the histogram of Colorado Rapids' salaries in Example 2 as being **right skewed**.

Exercise

We can think of the histogram of Colorado Rapids' salaries in Example 2 as being **right skewed**.

Which statistic do you think will be *larger*, the **mean** salary or the **median**?

Exercise

We can think of the histogram of Colorado Rapids' salaries in Example 2 as being **right skewed**.

Which statistic do you think will be *larger*, the **mean** salary or the **median**?

Which do you think is *more representative* of a typical player's salary?

Exercise

We can think of the histogram of Colorado Rapids' salaries in Example 2 as being **right skewed**.

Which statistic do you think will be *larger*, the **mean** salary or the **median**?

Which do you think is *more representative* of a typical player's salary?

(We saw in an earlier example that the mean is $157,086. It turns out that the median is $80,000).

Guidelines for Choosing a Measure of Center

- Here are suggestions for choosing between the mean and the median for reporting a typical value in a data set:
  - The mean is preferable for data that don't have any strong outliers and whose distribution is not too skewed.
  - The median the preferable if the data set has strong outliers or its distribution is very skewed (in either direction).